In *Proceedings of HLT/NAACL 2004*

Unsupervised Learning of Contextual Role Knowledge for Coreference Resolution

David Bean
Attensity Corporation, Suite 600
Gateway One 90 South 400 West
Salt Lake City, UT 84101
dbean@attensity.com

Ellen Riloff
School of Computing
University of Utah
Salt Lake City, UT 84112
riloff@cs.utah.edu

Abstract

We present a coreference resolver called BABAR that uses contextual role knowledge to evaluate possible antecedents for an anaphor. BABAR uses information extraction patterns to identify contextual roles and creates four contextual role knowledge sources using unsupervised learning. These knowledge sources determine whether the contexts surrounding an anaphor and antecedent are compatible. BABAR applies a Dempster-Shafer probabilistic model to make resolutions based on evidence from the contextual role knowledge sources as well as general knowledge sources. Experiments in two domains showed that the contextual role knowledge improved coreference performance, especially on pronouns.

1 Introduction

The problem of coreference resolution has received considerable attention, including theoretical discourse models (e.g., (Grosz et al., 1995; Grosz and Sidner, 1998)), syntactic algorithms (e.g., (Hobbs, 1978; Lappin and Leass, 1994)), and supervised machine learning systems (Aone and Bennett, 1995; McCarthy and Lehnert, 1995; Ng and Cardie, 2002; Soon et al., 2001). Most computational models for coreference resolution rely on properties of the anaphor and candidate antecedent, such as lexical matching, grammatical and syntactic features, semantic agreement, and positional information.

The focus of our work is on the use of *contextual role knowledge* for coreference resolution. A contextual role represents the role that a noun phrase plays in an event or relationship. Our work is motivated by the observation that contextual roles can be critically important in determining the referent of a noun phrase. Consider the following sentences:

(a) *Jose Maria Martinez, Roberto Lisandy, and Dino Rossy, who were staying at a Tecun Uman hotel, were kidnapped by armed men who took them to an unknown place.*

(b) *After **they** were released...*

(c) *After **they** blindfolded the men...*

In (b) "they" refers to the kidnapping victims, but in (c) "they" refers to the armed men. The role that each noun phrase plays in the kidnapping event is key to distinguishing these cases. The correct resolution in sentence (b) comes from knowledge that people who are kidnapped are often subsequently released. The correct resolution in sentence (c) depends on knowledge that kidnappers frequently blindfold their victims.

We have developed a coreference resolver called BABAR that uses contextual role knowledge to make coreference decisions. BABAR employs information extraction techniques to represent and learn role relationships. Each pattern represents the role that a noun phrase plays in the surrounding context. BABAR uses unsupervised learning to acquire this knowledge from plain text without the need for annotated training data. Training examples are generated automatically by identifying noun phrases that can be easily resolved with their antecedents using lexical and syntactic heuristics. BABAR then computes statistics over the training examples measuring the frequency with which extraction patterns and noun phrases co-occur in coreference resolutions.

In this paper, Section 2 begins by explaining how contextual role knowledge is represented and learned. Section 3 describes the complete coreference resolution model, which uses the contextual role knowledge as well as more traditional coreference features. Our coreference resolver also incorporates an existential noun phrase recognizer and a Dempster-Shafer probabilistic model to make resolution decisions. Section 4 presents experimen-

tal results on two corpora: the MUC-4 terrorism corpus, and Reuters texts about natural disasters. Our results show that BABAR achieves good performance in both domains, and that the contextual role knowledge improves performance, especially on pronouns. Finally, Section 5 explains how BABAR relates to previous work, and Section 6 summarizes our conclusions.

2 Learning Contextual Role Knowledge

In this section, we describe how contextual role knowledge is represented and learned. Section 2.1 describes how BABAR generates training examples to use in the learning process. We refer to this process as *Reliable Case Resolution* because it involves finding cases of anaphora that can be easily resolved with their antecedents. Section 2.2 then describes our representation for contextual roles and four types of contextual role knowledge that are learned from the training examples.

2.1 Reliable Case Resolutions

The first step in the learning process is to generate training examples consisting of anaphor/antecedent resolutions. BABAR uses two methods to identify anaphors that can be easily and reliably resolved with their antecedent: *lexical seeding* and *syntactic seeding*.

2.1.1 Lexical Seeding

It is generally not safe to assume that multiple occurrences of a noun phrase refer to the same entity. For example, *the company* may refer to Company X in one paragraph and Company Y in another. However, lexically similar NPs usually refer to the same entity in two cases: proper names and existential noun phrases. BABAR uses a named entity recognizer to identify proper names that refer to people and companies. Proper names are assumed to be coreferent if they match exactly, or if they closely match based on a few heuristics. For example, a person's full name will match with just their last name (e.g., "George Bush" and "Bush"), and a company name will match with and without a corporate suffix (e.g., "IBM Corp." and "IBM"). Proper names that match are resolved with each other.

The second case involves *existential* noun phrases (Allen, 1995), which are noun phrases that uniquely specify an object or concept and therefore do not need a prior referent in the discourse. In previous work (Bean and Riloff, 1999), we developed an unsupervised learning algorithm that automatically recognizes definite NPs that are existential without syntactic modification because their meaning is universally understood. For example, a story can mention "the FBI", "the White House", or "the weather" without any prior referent in the story.

Although these existential NPs do not need a prior referent, they may occur multiple times in a document. By definition, each existential NP uniquely specifies an object or concept, so we can infer that all instances of the same existential NP are coreferent (e.g., "the FBI" always refers to the same entity). Using this heuristic, BABAR identifies existential definite NPs in the training corpus using our previous learning algorithm (Bean and Riloff, 1999) and resolves all occurrences of the same existential NP with each another.[1]

2.1.2 Syntactic Seeding

BABAR also uses syntactic heuristics to identify anaphors and antecedents that can be easily resolved. Table 1 briefly describes the seven syntactic heuristics used by BABAR to resolve noun phrases. Words and punctuation that appear in brackets are considered optional. The anaphor and antecedent appear in boldface.

1. Reflexive pronouns with only 1 NP in scope. Ex: ***The regime*** *gives* ***itself*** *the right...*
2. Relative pronouns with only 1 NP in scope. Ex: *The* ***brigade, which*** *attacked ...*
3. Some cases of the pattern 'NP to-be NP". Ex: ***Mr. Cristiani*** *is* ***the president*** *...*
4. Some cases of 'NP said [that] it/they" Ex: ***The government*** *said* ***it*** *...*
5. Some cases of "[Locative-prep] NP [,] where" Ex: *He was found in* ***San Jose, where*** *...*
6. Simple appositives of the form "NP, NP" Ex: ***Mr. Cristiani, president*** *of the country ...*
7. PPs containing "by" and a gerund followed by "it" Ex: *Mr. Bush disclosed* ***the policy*** *by reading* ***it****...*

Table 1: Syntactic Seeding Heuristics

BABAR's reliable case resolution heuristics produced a substantial set of anaphor/antecedent resolutions that will be the training data used to learn contextual role knowledge. For terrorism, BABAR generated 5,078 resolutions: 2,386 from lexical seeding and 2,692 from syntactic seeding. For natural disasters, BABAR generated 20,479 resolutions: 11,652 from lexical seeding and 8,827 from syntactic seeding.

2.2 Contextual Role Knowledge

Our representation of contextual roles is based on information extraction patterns that are converted into simple caseframes. First, we describe how the caseframes are represented and learned. Next, we describe four contextual role knowledge sources that are created from the training examples and the caseframes.

2.2.1 The Caseframe Representation

Information extraction (IE) systems use extraction patterns to identify noun phrases that play a specific role in

[1] Our implementation only resolves NPs that occur in the same document, but in retrospect, one could probably resolve instances of the same existential NP in different documents too.

an event. For IE, the system must be able to distinguish between semantically similar noun phrases that play different roles in an event. For example, management succession systems must distinguish between a person who is fired and a person who is hired. Terrorism systems must distinguish between people who perpetrate a crime and people who are victims of a crime.

We applied the AutoSlog system (Riloff, 1996) to our unannotated training texts to generate a set of extraction patterns for each domain. Each extraction pattern represents a linguistic expression and a syntactic position indicating where a role filler can be found. For example, kidnapping victims should be extracted from the subject of the verb "kidnapped" when it occurs in the passive voice (the short-hand representation of this pattern would be *"<subject> were kidnapped"*). The types of patterns produced by AutoSlog are outlined in (Riloff, 1996).

Ideally we'd like to know the thematic role of each extracted noun phrase, but AutoSlog does not generate thematic roles. As a (crude) approximation, we normalize the extraction patterns with respect to active and passive voice and label those extractions as agents or patients. For example, the passive voice pattern *"<subject> were kidnapped"* and the active voice pattern *"kidnapped <direct_object>"* are merged into a single normalized pattern *"kidnapped <patient>"*.[2] For the sake of simplicity, we will refer to these normalized extraction patterns as *caseframes*.[3] These caseframes can capture two types of contextual role information: (1) thematic roles corresponding to events (e.g, *"<agent> kidnapped" or "kidnapped <patient>"*), and (2) predicate-argument relations associated with both verbs and nouns (e.g., *"kidnapped for <np>"* or *"vehicle with <np>"*).

We generate these caseframes automatically by running AutoSlog over the training corpus exhaustively so that it literally generates a pattern to extract every noun phrase in the corpus. The learned patterns are then normalized and applied to the corpus. This process produces a large set of caseframes coupled with a list of the noun phrases that they extracted. The contextual role knowledge that BABAR uses for coreference resolution is derived from this caseframe data.

2.2.2 The Caseframe Network

The first type of contextual role knowledge that BABAR learns is the **Caseframe Network (CFNet)**, which identifies caseframes that co-occur in anaphor/antecedent resolutions. Our assumption is that caseframes that co-occur in resolutions often have a conceptual relationship in the discourse. For example, co-occurring caseframes may reflect synonymy (e.g., *"<patient> kidnapped"* and *"<patient> abducted"*) or related events (e.g., *"<patient> kidnapped"* and *"<patient> released"*). We do not attempt to identify the types of relationships that are found. BABAR merely identifies caseframes that frequently co-occur in coreference resolutions.

Terrorism	Natural Disasters
murder of <NP> killed <patient>	<agent> damaged was injured in <NP>
<agent> reported <agent> added	<agent> occurred cause of <NP>
<agent> stated <agent> added	<agent> wreaked <agent> crossed
perpetrated <patient> condemned <patient>	driver of <NP> <agent> carrying

Figure 1: Caseframe Network Examples

Figure 1 shows examples of caseframes that co-occur in resolutions, both in the terrorism and natural disaster domains. The terrorism examples reflect fairly obvious relationships: people who are murdered are killed; agents that "report" things also "add" and "state" things; crimes that are "perpetrated" are often later "condemned". In the natural disasters domain, agents are often forces of nature, such as hurricanes or wildfires. Figure 1 reveals that an event that "damaged" objects may also cause injuries; a disaster that "occurred" may be investigated to find its "cause"; a disaster may "wreak" havoc as it "crosses" geographic regions; and vehicles that have a "driver" may also "carry" items.

During coreference resolution, the caseframe network provides evidence that an anaphor and prior noun phrase might be coreferent. Given an anaphor, BABAR identifies the caseframe that would extract it from its sentence. For each candidate antecedent, BABAR identifies the caseframe that would extract the candidate, pairs it with the anaphor's caseframe, and consults the CF Network to see if this pair of caseframes has co-occurred in previous resolutions. If so, the CF Network reports that the anaphor and candidate may be coreferent.

2.2.3 Lexical Caseframe Expectations

The second type of contextual role knowledge learned by BABAR is *Lexical Caseframe Expectations*, which are used by the **CFLex** knowledge source. For each caseframe, BABAR collects the head nouns of noun phrases that were extracted by the caseframe in the training corpus. For each resolution in the training data, BABAR also associates the co-referring expression of an NP with the NP's caseframe. For example, if X and Y are coreferent, then both X and Y are considered to co-occur with the caseframe that extracts X as well as the caseframe that

[2]This normalization is performed syntactically without semantics, so the agent and patient roles are not guaranteed to hold, but they usually do in practice.

[3]These are not full case frames in the traditional sense, but they approximate a simple case frame with a single slot.

extracts Y. We will refer to the set of nouns that co-occur with a caseframe as the *lexical expectations* of the caseframe. Figure 2 shows examples of lexical expectations that were learned for both domains.

Terrorism
Caseframe: *engaged in <NP>* **NPs:** activity, battle, clash, dialogue, effort, fight, group, shoot-out, struggle, village, violence
Caseframe: *ambushed <patient>* **NPs:** company, convoy, helicopter, member, motorcade, move, Ormeno, patrol, position, response, soldier, they, troops, truck, vehicle, which
Natural Disasters
Caseframe: *battled through <NP>* **NPs:** flame, night, smoke, wall
Caseframe: *braced for <NP>* **NPs:** arrival, battering, catastrophe, crest, Dolly, epidemics, evacuate, evacuation, flood, flooding, front, Hortense, hurricane, misery, rains, river, storm, surge, test, typhoon.

Figure 2: Lexical Caseframe Expectations

To illustrate how lexical expectations are used, suppose we want to determine whether noun phrase X is the antecedent for noun phrase Y. If they are coreferent, then X and Y should be substitutable for one another in the story.[4] Consider these sentences:

(S1) *Fred was killed by a masked man with* ***a revolver****.*
(S2) *The burglar fired* ***the gun*** *three times and fled.*

"The gun" will be extracted by the caseframe *"fired <patient>"*. Its correct antecedent is "a revolver", which is extracted by the caseframe *"killed with <NP>"*. If "gun" and "revolver" refer to the same object, then it should also be acceptable to say that Fred was *"killed with a gun"* and that the burglar *"fired a revolver"*.

During coreference resolution, BABAR checks (1) whether the anaphor is among the lexical expectations for the caseframe that extracts the candidate antecedent, and (2) whether the candidate is among the lexical expectations for the caseframe that extracts the anaphor. If either case is true, then CFLex reports that the anaphor and candidate might be coreferent.

2.2.4 Semantic Caseframe Expectations

The third type of contextual role knowledge learned by BABAR is *Semantic Caseframe Expectations*. Semantic expectations are analogous to lexical expectations except that they represent semantic classes rather than nouns. For each caseframe, BABAR collects the semantic classes associated with the head nouns of NPs that were extracted by the caseframe. As with lexical expections, the semantic classes of co-referring expressions are collected too. We will refer to the semantic classes that co-occur with a caseframe as the *semantic expectations* of the caseframe. Figure 3 shows examples of semantic expectations that were learned. For example, BABAR learned that agents that "assassinate" or "investigate a cause" are usually *humans* or *groups* (i.e., organizations).

Terrorism	
Caseframe	**Semantic Classes**
<agent> assassinated	*group, human*
investigation into <NP>	*event*
exploded outside <NP>	*building*
Natural Disasters	
Caseframe	**Semantic Classes**
<agent> investigating cause	*group, human*
survivor of <NP>	*event, natphenom*
hit with <NP>	*attribute, natphenom*

Figure 3: Semantic Caseframe Expectations

For each domain, we created a semantic dictionary by doing two things. First, we parsed the training corpus, collected all the noun phrases, and looked up each head noun in WordNet (Miller, 1990). We tagged each noun with the top-level semantic classes assigned to it in WordNet. Second, we identified the 100 most frequent nouns in the training corpus and manually labeled them with semantic tags. This step ensures that the most frequent terms for each domain are labeled (in case some of them are not in WordNet) and labeled with the sense most appropriate for the domain.

Initially, we planned to compare the semantic classes of an anaphor and a candidate and infer that they might be coreferent if their semantic classes intersected. However, using the top-level semantic classes of WordNet proved to be problematic because the class distinctions are too coarse. For example, both a chair and a truck would be labeled as *artifacts*, but this does not at all suggest that they are coreferent. So we decided to use semantic class information only to rule out candidates. If two nouns have mutually exclusive semantic classes, then they cannot be coreferent. This solution also obviates the need to perform word sense disambiguation. Each word is simply tagged with the semantic classes corresponding to all of its senses. If these sets do not overlap, then the words cannot be coreferent.

The semantic caseframe expectations are used in two ways. One knowledge source, called **WordSem-CFSem**, is analogous to CFLex: it checks whether the anaphor and candidate antecedent are substitutable for one another, but based on their semantic classes instead of the words themselves. Given an anaphor and candidate, BABAR checks (1) whether the semantic classes of the anaphor intersect with the semantic expectations of the caseframe that extracts the candidate, and (2) whether the semantic classes of the candidate intersect with the semantic ex-

[4] They may not be perfectly substitutable, for example one NP may be more specific (e.g., 'he" vs. "John F. Kennedy"). But in most cases they can be used interchangably.

pectations of the caseframe that extracts the anaphor. If one of these checks fails then this knowledge source reports that the candidate is not a viable antecedent for the anaphor.

A different knowledge source, called **CFSem-CFSem**, compares the semantic expectations of the caseframe that extracts the anaphor with the semantic expectations of the caseframe that extracts the candidate. If the semantic expectations do not intersect, then we know that the caseframes extract mutually exclusive types of noun phrases. In this case, this knowledge source reports that the candidate is not a viable antecedent for the anaphor.

2.3 Assigning Evidence Values

Contextual role knowledge provides evidence as to whether a candidate is a plausible antecedent for an anaphor. The two knowledge sources that use semantic expectations, WordSem-CFSem and CFSem-CFSem, always return values of -1 or 0. -1 means that an NP should be ruled out as a possible antecedent, and 0 means that the knowledge source remains neutral (i.e., it has no reason to believe that they cannot be coreferent).

The CFLex and CFNet knowledge sources provide positive evidence that a candidate NP and anaphor might be coreferent. They return a value in the range [0,1], where 0 indicates neutrality and 1 indicates the strongest belief that the candidate and anaphor are coreferent. BABAR uses the log-likelihood statistic (Dunning, 1993) to evaluate the strength of a co-occurrence relationship. For each co-occurrence relation (noun/caseframe for CFLex, and caseframe/caseframe for CFNet), BABAR computes its log-likelihood value and looks it up in the χ^2 table to obtain a confidence level. The confidence level is then used as the belief value for the knowledge source. For example, if CFLex determines that the log-likelihood statistic for the co-occurrence of a particular noun and caseframe corresponds to the 90% confidence level, then CFLex returns .90 as its belief that the anaphor and candidate are coreferent.

3 The Coreference Resolution Model

Given a document to process, BABAR uses four modules to perform coreference resolution. First, a *non-anaphoric NP classifier* identifies definite noun phrases that are existential, using both syntactic rules and our learned existential NP recognizer (Bean and Riloff, 1999), and removes them from the resolution process. Second, BABAR performs reliable case resolution to identify anaphora that can be easily resolved using the lexical and syntactic heuristics described in Section 2.1. Third, all remaining anaphora are evaluated by 11 different knowledge sources: the four contextual role knowledge sources just described and seven general knowledge sources. Finally, a Dempster-Shafer probabilistic model evaluates the evidence provided by the knowledge sources for all candidate antecedents and makes the final resolution decision. In this section, we describe the seven general knowledge sources and explain how the Dempster-Shafer model makes resolutions.

3.1 General Knowledge Sources

Figure 4 shows the seven general knowledge sources (KSs) that represent features commonly used for coreference resolution. The gender, number, and scoping KSs eliminate candidates from consideration. The scoping heuristics are based on the anaphor type: for reflexive pronouns the scope is the current clause, for relative pronouns it is the prior clause following its VP, for personal pronouns it is the anaphor's sentence and two preceding sentences, and for definite NPs it is the anaphor's sentence and eight preceding sentences. The semantic agreement KS eliminates some candidates, but also provides positive evidence in one case: if the candidate and anaphor both have semantic tags *human*, *company*, *date*, or *location* that were assigned via NER or the manually labeled dictionary entries. The rationale for treating these semantic labels differently is that they are specific and reliable (as opposed to the WordNet classes, which are more coarse and more noisy due to polysemy).

KS	Function
Gender	filters candidate if gender doesn't agree.
Number	filters candidate if number doesn't agree.
Scoping	filters candidate if outside the anaphor's scope.
Semantic	(a) filters candidate if its semantic tags don't intersect with those of the anaphor. (b) supports candidate if selected semantic tags match those of the anaphor.
Lexical	computes degree of lexical overlap between the candidate and the anaphor.
Recency	computes the relative distance between the candidate and the anaphor.
SynRole	computes relative frequency with which the candidate's syntactic role occurs in resolutions.

Figure 4: General Knowledge Sources

The Lexical KS returns 1 if the candidate and anaphor are identical, 0.5 if their head nouns match, and 0 otherwise. The Recency KS computes the distance between the candidate and the anaphor relative to its scope. The SynRole KS computes the relative frequency with which the candidates' syntactic role (subject, direct object, PP object) appeared in resolutions in the training set. During development, we sensed that the Recency and Synrole KSs did not deserve to be on equal footing with the other KSs because their knowledge was so general. Consequently, we cut their evidence values in half to lessen their influence.

3.2 The Dempster-Shafer Decision Model

BABAR uses a Dempster-Shafer decision model (Stefik, 1995) to combine the evidence provided by the knowledge sources. Our motivation for using Dempster-Shafer is that it provides a well-principled framework for combining evidence from multiple sources with respect to competing hypotheses. In our situation, the competing hypotheses are the possible antecedents for an anaphor.

An important aspect of the Dempster-Shafer model is that it operates on sets of hypotheses. If evidence indicates that hypotheses C and D are less likely than hypotheses A and B, then probabilities are redistributed to reflect the fact that {A, B} is more likely to contain the answer than {C, D}. The ability to redistribute belief values across sets rather than individual hypotheses is key. The evidence may not say anything about whether A is more likely than B, only that C and D are not likely.

Each set is assigned two values: *belief* and *plausibility*. Initially, the Dempster-Shafer model assumes that all hypotheses are equally likely, so it creates a set called θ that includes all hypotheses. θ has a belief value of 1.0, indicating complete certainty that the correct hypothesis is included in the set, and a plausibility value of 1.0, indicating that there is no evidence for competing hypotheses.[5] As evidence is collected and the likely hypotheses are whittled down, belief is redistributed to subsets of θ.

Formally, the Dempster-Shafer theory defines a probability density function $m(S)$, where S is a set of hypotheses. $m(S)$ represents the belief that the correct hypothesis is included in S. The model assumes that evidence also arrives as a probability density function (pdf) over sets of hypotheses.[6] Integrating new evidence into the existing model is therefore simply a matter of defining a function to merge pdfs, one representing the current belief system and one representing the beliefs of the new evidence. The Dempster-Shafer rule for combining pdfs is:

$$m_3(S) = \frac{\sum_{X \cap Y = S} m_1(X) * m_2(Y)}{1 - \sum_{X \cap Y = \emptyset} m_1(X) * m_2(Y)} \quad (1)$$

All sets of hypotheses (and their corresponding belief values) in the current model are crossed with the sets of hypotheses (and belief values) provided by the new evidence. Sometimes, however, these beliefs can be contradictory. For example, suppose the current model assigns a belief value of .60 to {A, B}, meaning that it is 60% sure that the correct hypothesis is either A or B. Then new evidence arrives with a belief value of .70 assigned to {C}, meaning that it is 70% sure the correct hypothesis is C. The intersection of these sets is the null set because these beliefs are contradictory. The belief value that would have been assigned to the intersection of these sets is .60*.70=.42, but this belief has nowhere to go because the null set is not permissible in the model.[7] So this probability mass (.42) has to be redistributed. Dempster-Shafer handles this by re-normalizing all the belief values with respect to only the non-null sets (this is the purpose of the denominator in Equation 1).

In our coreference resolver, we define θ to be the set of all candidate antecedents for an anaphor. Each knowledge source then assigns a probability estimate to each candidate, which represents its belief that the candidate is the antecedent for the anaphor. The probabilities are incorporated into the Dempster-Shafer model using Equation 1. To resolve the anaphor, we survey the final belief values assigned to each candidate's singleton set. If a candidate has a belief value $\geq .50$, then we select that candidate as the antecedent for the anaphor. If no candidate satisfies this condition (which is often the case), then the anaphor is left unresolved. One of the strengths of the Dempster-Shafer model is its natural ability to recognize when several credible hypotheses are still in play. In this situation, BABAR takes the conservative approach and declines to make a resolution.

4 Evaluation Results

4.1 Corpora

We evaluated BABAR on two domains: terrorism and natural disasters. We used the MUC-4 terrorism corpus (MUC-4 Proceedings, 1992) and news articles from the Reuter's text collection[8] that had a subject code corresponding to natural disasters. For each domain, we created a blind test set by manually annotating 40 documents with anaphoric chains, which represent sets of noun phrases that are coreferent (as done for MUC-6 (MUC-6 Proceedings, 1995)). In the terrorism domain, 1600 texts were used for training and the 40 test documents contained 322 anaphoric links. For the disasters domain, 8245 texts were used for training and the 40 test documents contained 447 anaphoric links.

In recent years, coreference resolvers have been evaluated as part of MUC-6 and MUC-7 (MUC-7 Proceedings, 1998). We considered using the MUC-6 and MUC-7 data sets, but their training sets were far too small to learn reliable co-occurrence statistics for a large set of contextual role relationships. Therefore we opted to use the much

[5] Initially there are no competing hypotheses because all hypotheses are included in θ by definition.

[6] Our knowledge sources return some sort of probability estimate, although in some cases this estimate is not especially well-principled (e.g., the Recency KS).

[7] The Dempster-Shafer theory assumes that one of the hypotheses in θ is correct, so eliminating all of the hypotheses violates this assumption.

[8] Volume 1, English language, 1996-1997, Format version 1, correction level 0

Anaphor	Terrorism Rec	Pr	F	Disasters Rec	Pr	F
Def. NPs	.43	.79	.55	.42	.91	.58
Pronouns	.50	.72	.59	.42	.82	.56
Total	.46	.76	.57	.42	.87	.57

Table 2: General Knowledge Sources

Anaphor	Terrorism Rec	Pr	F	Disasters Rec	Pr	F
Def. NPs	.45	.71	.55	.46	.84	.59
Pronouns	.63	.73	.68	.57	.79	.66
Total	.53	.73	.61	.51	.82	.63

Table 3: General + Contextual Role Knowledge Sources

larger MUC-4 and Reuters corpora.[9]

4.2 Experiments

We adopted the MUC-6 guidelines for evaluating coreference relationships based on transitivity in anaphoric chains. For example, if $\{NP_1, NP_2, NP_3\}$ are all coreferent, then each NP must be linked to one of the other two NPs. First, we evaluated BABAR using only the seven general knowledge sources. Table 2 shows BABAR's performance. We measured recall (Rec), precision (Pr), and the F-measure (F) with recall and precision equally weighted. BABAR achieved recall in the 42-50% range for both domains, with 76% precision overall for terrorism and 87% precision for natural disasters. We suspect that the higher precision in the disasters domain may be due to its substantially larger training corpus.

Table 3 shows BABAR's performance when the four contextual role knowledge sources are added. The F-measure score increased for both domains, reflecting a substantial increase in recall with a small decrease in precision. The contextual role knowledge had the greatest impact on pronouns: +13% recall for terrorism and +15% recall for disasters, with a +1% precision gain in terrorism and a small precision drop of -3% in disasters.

The difference in performance between pronouns and definite noun phrases surprised us. Analysis of the data revealed that the contextual role knowledge is especially helpful for resolving pronouns because, in general, they are semantically weaker than definite NPs. Since pronouns carry little semantics of their own, resolving them depends almost entirely on context. In contrast, even though context can be helpful for resolving definite NPs, context can be trumped by the semantics of the nouns themselves. For example, even if the contexts surrounding an anaphor and candidate match exactly, they are not coreferent if they have substantially different meanings (e.g., "the mayor" vs. "the journalist").

[9]We would be happy to make our manually annotated test data available to others who also want to evaluate their coreference resolver on the MUC-4 or Reuters collections.

	Pronouns Rec	Pr	F	Definite NPs Rec	Pr	F
No CF KSs	.50	.72	.59	.43	.79	.55
CFLex	.56	.74	.64	.42	.73	.53
CFNet	.56	.74	.64	.43	.74	.54
CFSem-CFSem	.58	.76	.66	.44	.76	.56
WordSem-CFSem	.61	.74	.67	.45	.76	.56
All CF KSs	.63	.73	.68	.45	.71	.55

Table 4: Individual Performance of KSs for Terrorism

	Pronouns Rec	Pr	F	Definite NPs Rec	Pr	F
No CF KSs	.42	.82	.56	.42	.91	.58
CFLex	.48	.83	.61	.44	.88	.59
CFNet	.45	.82	.58	.43	.88	.57
CFSem-CFSem	.51	.81	.62	.44	.87	.58
WordSem-CFSem	.52	.79	.63	.43	.86	.57
All CF KSs	.57	.79	.66	.46	.84	.59

Table 5: Individual Performance of KSs for Disasters

We also performed experiments to evaluate the impact of each type of contextual role knowledge separately. Tables 4 and 5 show BABAR's performance when just one contextual role knowledge source is used at a time. For definite NPs, the results are a mixed bag: some knowledge sources increased recall a little, but at the expense of some precision. For pronouns, however, all of the knowledge sources increased recall, often substantially, and with little if any decrease in precision. This result suggests that all of contextual role KSs can provide useful information for resolving anaphora. Tables 4 and 5 also show that putting all of the contextual role KSs in play at the same time produces the greatest performance gain. There are two possible reasons: (1) the knowledge sources are resolving different cases of anaphora, and (2) the knowledge sources provide multiple pieces of evidence in support of (or against) a candidate, thereby acting synergistically to push the Dempster-Shafer model over the belief threshold in favor of a single candidate.

5 Related Work

Many researchers have developed coreference resolvers, so we will only discuss the methods that are most closely related to BABAR. Dagan and Itai (Dagan and Itai, 1990) experimented with co-occurrence statistics that are similar to our lexical caseframe expectations. Their work used subject-verb, verb-object, and adjective-noun relations to compare the contexts surrounding an anaphor and candidate. However their work did not consider other types of lexical expectations (e.g., PP arguments), semantic expectations, or context comparisons like our caseframe network.

(Niyu et al., 1998) used unsupervised learning to ac-

quire gender, number, and animacy information from resolutions produced by a statistical pronoun resolver. The learned information was recycled back into the resolver to improve its performance. This approach is similar to BABAR in that they both acquire knowledge from earlier resolutions. (Kehler, 1997) also used a Dempster-Shafer model to merge evidence from different sources for template-level coreference.

Several coreference resolvers have used supervised learning techniques, such as decision trees and rule learners (Aone and Bennett, 1995; McCarthy and Lehnert, 1995; Ng and Cardie, 2002; Soon et al., 2001). These systems rely on a training corpus that has been manually annotated with coreference links.

6 Conclusions

The goal of our research was to explore the use of contextual role knowledge for coreference resolution. We identified three ways that contextual roles can be exploited: (1) by identifying caseframes that co-occur in resolutions, (2) by identifying nouns that co-occur with caseframes and using them to cross-check anaphor/candidate compatibility, (3) by identifying semantic classes that co-occur with caseframes and using them to cross-check anaphor/candidate compatability. We combined evidence from four contextual role knowledge sources with evidence from seven general knowledge sources using a Dempster-Shafer probabilistic model.

Our coreference resolver performed well in two domains, and experiments showed that each contextual role knowledge source contributed valuable information. We found that contextual role knowledge was more beneficial for pronouns than for definite noun phrases. This suggests that different types of anaphora may warrant different treatment: definite NP resolution may depend more on lexical semantics, while pronoun resolution may depend more on contextual semantics. In future work, we plan to follow-up on this approach and investigate other ways that contextual role knowledge can be used.

7 Acknowledgements

This work was supported in part by the National Science Foundation under grant IRI-9704240. The inventions disclosed herein are the subject of a patent application owned by the University of Utah and licensed on an exclusive basis to Attensity Corporation.

References

J. Allen. 1995. *Natural Language Understanding*. Benjamin/Cummings Press, Redwood City, CA.

C. Aone and S. Bennett. 1995. Applying Machine Learning to Anaphora Resolution. In *IJCAI-95 Workshop on New Approaches to Learning for NLP*.

D. Bean and E. Riloff. 1999. Corpus-Based Identification of Non-Anaphoric Noun Phrases. In *Proc. of the 37th Annual Meeting of the Association for Computational Linguistics*.

I. Dagan and A. Itai. 1990. Automatic Processing of Large Corpora for the Resolution of Anaphora References. In *Proceedings of the Thirteenth International Conference on Computational Linguistics (COLING-90)*, pages 330–332.

T. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

B. Grosz and C. Sidner. 1998. Lost Intuitions and Forgotten Intentions. In M. Walker, A. Joshi, and E. Prince, editors, *Centering Theory in Discourse*, pages 89–112. Clarendon Press.

B. Grosz, A. Joshi, and S. Weinstein. 1995. Centering: A Framework for Modeling the Local Coherence of Discourse. *Computational Linguistics*, 21(2):203–226.

J. Hobbs. 1978. Resolving Pronoun References. *Lingua*, 44(4):311–338.

A. Kehler. 1997. Probabilistic Coreference in Information Extraction. In *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing*.

S. Lappin and H. Leass. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics*, 20(4):535–561.

J. McCarthy and W. Lehnert. 1995. Using Decision Trees for Coreference Resolution. In *Proc. of the Fourteenth International Joint Conference on Artificial Intelligence*.

G. Miller. 1990. Wordnet: An On-line Lexical Database. *International Journal of Lexicography*, 3(4).

MUC-4 Proceedings. 1992. *Proceedings of the Fourth Message Understanding Conference (MUC-4)*.

MUC-6 Proceedings. 1995. *Proceedings of the Sixth Message Understanding Conference (MUC-6)*.

MUC-7 Proceedings. 1998. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*.

V. Ng and C. Cardie. 2002. Improving Machine Learning Approaches to Coreference Resolution. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

G. Niyu, J. Hale, and E. Charniak. 1998. A statistical approach to anaphora resolution. In *Proceedings of the Sixth Workshop on Very Large Corpora*.

E. Riloff. 1996. An Empirical Study of Automated Dictionary Construction for Information Extraction in Three Domains. *Artificial Intelligence*, 85:101–134.

W. Soon, H. Ng, and D. Lim. 2001. A Machine Learning Approach to Coreference of Noun Phrases. *Computational Linguistics*, 27(4):521–541.

M. Stefik. 1995. *Introduction to Knowledge Systems*. Morgan Kaufmann, San Francisco, CA.